

# Predicting function from sequence in a large multifunctional toxin family

Malhotra, A.; Creer, S.; Harris, J.B.; Stöcklin, R.; Favereau, P.; Thorpe, R.S.

**Toxicon**

*DOI:*
10.1016/j.toxicon.2013.06.019

Published: 03/07/2013

Publisher's PDF, also known as Version of record

Toxicon 72 (2013) 113–125

Contents lists available at SciVerse ScienceDirect

Toxicon

journal homepage: www.elsevier.com/locate/toxicon

# Predicting function from sequence in a large multifunctional toxin family

Anita Malhotra [a,*], Simon Creer [a], John B. Harris [b], Reto Stöcklin [c], Philippe Favreau [c], Roger S. Thorpe [a]

[a] School of Biological Sciences, College of Natural Sciences, Bangor University, Bangor LL57 2UW, UK
[b] Medical Toxicology Centre, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK
[c] Atheris Laboratories, Case Postale 314, CH-1233 Bernex-Geneva, Switzerland

ARTICLE INFO

*Article history:*
Received 4 April 2013
Received in revised form 21 June 2013
Accepted 26 June 2013
Available online 4 July 2013

*Keywords:*
Snake venom phospholipase $A_2$
Crotalinae
Discriminant function analysis
Sequence profiles
Functional prediction

ABSTRACT

Venoms contain active substances with highly specific physiological effects and are increasingly being used as sources of novel diagnostic, research and treatment tools for human disease. Experimental characterisation of individual toxin activities is a severe rate-limiting step in the discovery process, and *in-silico* tools which allow function to be predicted from sequence information are essential. Toxins are typically members of large multifunctional families of structurally similar proteins that can have different biological activities, and minor sequence divergence can have significant consequences. Thus, existing predictive tools tend to have low accuracy. We investigated a classification model based on physico-chemical attributes that can easily be calculated from amino-acid sequences, using over 250 (mostly novel) viperid phospholipase $A_2$ toxins. We also clustered proteins by sequence profiles, and carried out *in-vitro* tests for four major activities on a selection of isolated novel toxins, or crude venoms known to contain them. The majority of detected activities were consistent with predictions, in contrast to poor performance of a number of tested existing predictive methods. Our results provide a framework for comparison of active sites among different functional sub-groups of toxins that will allow a more targeted approach for identification of potential drug leads in the future.


## 1. Introduction

Animal toxins often form functionally diverse families, being based on a relatively limited number of basic scaffolds yet achieving a diverse range of physiological effects through interaction with a multitude of molecular targets. They offer a virtually unlimited pool of bioactive compounds with therapeutic and pharmacological potential, a fact which is attracting increasing interest in academic, industrial and medical arenas (King, 2011). Pre-screening of newly identified compounds with *in-silico* techniques to identify functional hypotheses for subsequent experimental testing is highly desirable but limited by current levels of accuracy of many existing bioinformatics methods (Clark and Radivojac, 2010; Koonin, 2000). Even computationally quite complex methods may have prediction accuracies of less than 50% when applied to functionally diverse protein families (Engelhardt et al., 2011). An excellent example is provided by toxins based

*Abbreviations used:* $PLA_2$, phospholipase $A_2$; $svPLA_2$, snake venom phospholipase $A_2$; DFA, discriminant function analysis; MW, molecular weight; MALDI–TOF, matrix-assisted laser-desorption ionisation–time-of-flight; LC–ES, liquid chromatography–electrospray ionisation tandem.

* Corresponding author. School of Biological Sciences, College of Natural Sciences, Bangor University, 3rd Floor ECW, Deiniol Road, Bangor LL57 2UW, UK. Tel.: +44 (0)1248 383735; fax: +44 (0)1248 370731.
*E-mail address:* a.malhotra@bangor.ac.uk (A. Malhotra).

0041-0101 
http://dx.doi.org/10.1016/j.toxicon.2013.06.019

on the phospholipase $A_2$ ($PLA_2$) enzyme scaffold, a major component of reptile venoms, which hydrolyse phospholipids to release lysophospholipids and fatty acids (Kini, 1997). They also have toxic activities (including pre- and, more rarely, post-synaptic neurotoxicity, myotoxicity, cardiotoxicity, anticoagulant and haemolytic activity) that are independent of the catalytic activity of the enzyme and many $PLA_2$ toxins are in fact phospholipase homologues, in which mutational changes to the active site have abolished the phospholipase activity. Toxicity can occur through highly specific direct binding to membrane-bound, intracellular receptors or coagulation factors present in mammalian blood, or through interactions dependant on the three-dimensional structure of the folded protein, either in monomeric or dimeric form (Chioato and Ward, 2003). Group II $PLA_2$s (most similar to non-toxic $PLA_2$s in mammalian synovial fluid and testes [Doley et al., 2009]) are especially significant in viperid snakes, where they may make up to 70% of the protein content of crude venom. They are frequently present as multiple isoforms in the venom of single species (Calvete et al., 2011), and even a single individual (Danse et al., 1997; Ogawa et al., 1992), and have been shown to be the most variable of all major protein families in the venom, both intra- and interspecifically (Sanz et al., 2006).

The proliferation of functional activity appears to be dependent on the mutation of highly specific surface residues, which are hypothesised to change the specific target of the protein and thus confer a new activity (Doley et al., 2009). Predictions have been made about the position of pharmacological sites following functional studies on isoenzymes (Kini, 2006; Kini and Iwanaga, 1986a,b), while chemical modification, site-directed mutagenic mapping, use of monoclonal and polyclonal antibodies and analysis of inhibitor interactions have identified particular residues or segments of the $PLA_2$ molecules that are involved in different activities (Doley et al., 2009). A more recent and promising line of research uses biomimetic synthetic peptides to narrow down potential pharmacological sites (Lomonte et al., 2010). However, these studies often disagree and have generally failed to allow prediction of activity in other isoforms of unknown activity. In recent years, the discovery rate of new toxins has increased exponentially, as venom gland transcriptomic (EST) studies have resulted in the description of hundreds of new toxins (Boldrini-França et al., 2009; Durban et al., 2011; Junqueira-de-Azevedo and Ho, 2002; Kashima et al., 2004; Wagstaff and Harrison, 2006; Zhang et al., 2006). More recent application of next-generation sequencing technology (Chatrath et al., 2011; Jiang et al., 2011; Rodrigues et al., 2012) to transcriptomics will further accelerate this process, as will the increasing ability to directly access the genome through extended read length and targeted sequencing (Glenn, 2011). However, current methods for studying pharmacological activity are generally labour-intensive and the functional characterisation of these new toxins is unlikely to keep pace (not unique to toxins, as the majority of protein sequences in databases lack functional annotation). Computer-generated annotations have been shown to be highly inaccurate (Schnoes et al., 2009) mainly as a result of over-prediction (i.e., annotation to functions that are more specific than the available evidence supports, sometimes naively based on homology to primary structures). This is likely to be the case for most animal toxins, which often retain the ancestral non-toxic structural scaffold, while evolving diverse potent and highly specific toxic activities. In some cases, the substitution of a single amino acid is enough to change the selectivity for another target (Ohno et al., 1998). In the case of $PLA_2$ toxins, the ancestral phospholipase activity may be readily predicted while failing to predict the main biological activity of the protein in question. Thus, predicting the function of snake venom proteins based on a common scaffold presents a challenge to bioinformaticians interested in the analysis of protein sequence–function relationships in general. Solving this problem will have a number of beneficial outcomes as many of the activities of these proteins are of great utility as research tools and potential drugs (Koh et al., 2006), especially in neurological (Sun et al., 2004), anti-cancer (Bazaa et al., 2010; Lomonte et al., 2010), anti-viral (Fenard et al., 1999; Meenakshisundaram et al., 2009) and anti-inflammatory (Coulthard et al., 2011) research.

In this paper, we report a model-based analysis of the largest dataset of $PLA_2$ Group II toxins to date, comprising 251 protein sequences. Of these, 73 are novel sequences derived from a genome-based survey of $PLA_2$ genes in pitvipers (Viperidae: Crotalinae), including 16 species for which no $PLA_2$ sequences exist in databases. Most of the newly investigated species belong to the Asian *Trimeresurus* radiation (Malhotra and Thorpe, 2004), which have been relatively understudied by toxinologists (Gowda et al., 2006; Soogarun et al., 2008; Tan and Tan, 1989; Tan et al., 1989; Wang et al., 2005). We used two methods with different conceptual bases. The first was based on ordination of proteins with experimentally characterised functions, using biologically meaningful features derived from their physico-chemical characteristics in a discriminant function analysis (DFA). The second was based on primary sequence structure, measured by sequence profiles. DFA correctly classified between 62 and 86% of toxins with known physiological functions, with a good match between structural similarity and predicted function in the profile-based clustering method. In marked contrast, a number of alternative protein prediction methods failed to correctly identify more than the basic enzymatic function of the $PLA_2$ scaffold. *In-vitro* tests for four major activities showed that the activity of the majority was consistent with predicted functions. An advantage of the methods applied here is that they do not require specially-written software as all methods are already readily available in the public domain. If required, a bioinformatics pipeline to scale up these analyses to take advantage of high-throughput datasets from large-scale drug discovery programs could easily be constructed.

## 2. Experimental

### 2.1. Samples, cloning of $PLA_2$ genes and quality checking

Samples were collected between 1992 and 2002 as part of a systematic study on Asian pitvipers (Malhotra and Thorpe, 2004) and were in the form of blood samples,

ethanol-preserved scale clips, liver, or muscle tissue. We amplified $PLA_2$ genes directly from genomic extracts using conserved primers located in the untranslated regions of the $PLA_2$ genes, cloned individual PCR products, and sequenced multiple positive clones using the primers and procedures described previously (Dawson et al., 2010). Similar sequences from individual samples were grouped for detection of PCR errors and construction of consensus sequences, based on a statistically robust method of determining the probability of obtaining PCR artefacts (Dawson et al., 2010). However, we modified the acceptance criterion such that the minimum number of differences separating two sequences that had confirmed translation products in the venom (detected by proteomic analysis) was used set the acceptance threshold, if this was less than the threshold value determined by the cumulative binomial distribution. We applied a number of methods for the detection of recombinant sequences in an alignment: RDP, Geneconv, Chimaera, 3SEQ (all implemented in RDP3 [Martin et al., 2005]). Those showing clear evidence of recombination within sequences derived from single individuals were removed as likely PCR artefacts. Remaining sequences were aligned by eye into exons and introns using known splice sites in a reference $PLA_2$ sequence from *Protobothrops flavoviridis* (D13383). The putative protein-coding sequence was assembled and translated using EXPASY tools (web.expasy.org). The UniProt database and literature sources were searched for additional non-redundant crotaline $PLA_2$ protein sequences and resulting database aligned using MUSCLE (Edgar, 2004), implemented within Jalview (Waterhouse et al., 2009).

### 2.2. Proteomics: analysis of expressed $PLA_2$ toxins

Venom samples were obtained from a single milking of individual snakes, dried, and stored at −4 °C. Our goal was a) to characterise the expression profile of $PLA_2$ toxins in the crude venom, and b) to isolate several $PLA_2$s for activity testing (which was limited by the amount of crude venom available). Crude venom samples from 132 specimens of 29 species of Crotalinae were analysed by MALDI–TOF (matrix-assisted laser-desorption ionisation–time-of-flight) MS as described previously (Creer et al., 2003). Some later analyses were carried out using an Ultraflex™ TOF/TOF (Bruker Daltonics, Germany) with only minor modifications of the protocol. Calibrants used in the MALDI–TOF analyses were bovine insulin, ubiquitin I, cytochrome C, and myoglobin. Most samples were analysed at least twice, with some samples being analysed in each different set of analyses, which were carried out over a number of years. To check the reproducibility of the venom profile within individuals, we also analysed venom samples from captive individuals that had been collected monthly over the course of one year. A limited number of samples were also analysed using LC–ES (liquid chromatography–electrospray ionisation tandem) MS, to check the accuracy and reproducibility of results, as described previously (Creer et al., 2003). The mass range between 13 and 14.5 kDa was analysed using Data Explorer Version 3.5.0.0 (PerSeptive Biosystems). 'Major' peaks were defined as those with greater than 30% maximum intensity for MALDI–TOF analysis, while for LC–MS they corresponded to compounds exhibiting a UV absorption (214 nm) superior to 15% of the relative maximum intensity for LC–MS. In case of co-eluting proteins, the MS spectrum was taken into account and only the major representatives are considered as 'major' forms. 'Secondary' peaks were those with less than 30% maximum intensity for MALDI–TOF analysis, or those which correspond to compounds exhibiting a UV absorption (214 nm) inferior to 15% of the relative maximum intensity for LC–MS. Observed masses were subsequently grouped together if their masses were within the limits of the accuracy of the method used to determine them (i.e., within 10Da for two masses determined using MALDI–TOF, 2Da for those determined by LC–ES–MS, or 6Da for a mass determined by MALDI–TOF compared to one determined by LC–ES–MS). This procedure is conservative in that some $PLA_2$s with masses within the limits given above may result from different underlying sequences, but it minimises the chances of false discovery. TagIdent (EXPASY) was used to search UniprotKB/Swissprot for matches with individual sequenced isoforms.

Isoform content is particularly diverse and variable in the Chinese bamboo viper *Viridovipera stejnegeri* on the island of Taiwan (Creer et al., 2003). The distribution of high molecular weight versus low molecular weight isoforms is not random and appears to be correlated with diet. On the basis of the MALDI–TOF profiles, we chose one specimen (T61) from Green island, Taiwan (5.4 mg of dry crude venom), and two specimens (T221 and T224) with a similar venom profile from Fujian province, China (4.7 mg combined weight of dried venom), in which high and low molecular weight $PLA_2$s respectively formed the major components of the venom. The purification of the $PLA_2$s was carried out using Reverse-phase HPLC on 1 mg of crude venom. All the fractions were manually collected and a MALDI–TOF–MS analysis was performed in order to confirm the final mass of each fraction. Finally, the quantity and purity of each manually collected fraction was assessed by size exclusion chromatography.

### 2.3. Biological activities of venoms

Haemorrhagic activity was assessed by exposing blood vessels serving unhatched chick embryos to filter paper discs (2 mm diameter) loaded with fixed concentrations of venom samples in 0.9% w/v NaCl (44), using *Bothrops jararaca* venom as a positive control and 0.9% w/v NaCl alone as a negative control (Sells et al., 1998). Haemorrhagic activity was measured as the time taken for a haemorrhagic corona to appear around the disc, and the area of the corona after continuous contact with the disc for 2hr. Myotoxic and neurotoxic activity were assessed by incubating mouse soleus muscles at room temperature in oxygenated Liley's fluid for three hours in the presence of samples of venom or venom fractions at a fixed concentration of 10 $\mu g\ ml^{-1}$. At the end of the period of incubation, muscles were lightly fixed, cryoprotected, frozen in liquid N2 and sectioned at 6 μm (TS) and 10 μm (LS). For the assessment of myotoxicity, sections were stained with H & E and evidence of frank necrosis, hyper-contraction, and oedematous separation of necrotic muscle fibres (Harris

et al., 1975) was sought. For the assessment of neurotoxicity sections were labelled with a primary antibody for synaptophysin (a protein specific to synaptic vesicles) and a primary antibody for neurofilament (a protein specific to axons) and then to a secondary antibody conjugated to a fluorescent tag. Each section was counter-labelled with alpha-bungarotoxin conjugated to a fluorescent tag to identify the ACh receptors at the neuromuscular junction. Neurotoxicity was assessed by the absence of labelling for synaptophysin at the neuromuscular junction, or by abnormal labelling of neurofilament (Dixon and Harris, 1999; Prasarnpun et al., 2005). At least two muscles were used for each compound.

### 2.4. Discriminant analysis of protein physico-chemical properties

We used SMS (http://www.bioinformatics.org/sms2/protein_gravy.html) and Protparam (EXPASY) to calculate a number of sequence-based features including pI (isoelectric point), MW (theoretical average molecular weight, without any correction made for disulphide bridges), net charge, GRAVY (GRand AVerage of hYdropathy [Kyte and Doolittle, 1982]), aliphatic index (a measure of the thermostability of globular proteins), instability index and amino acid composition (%). The 20 amino-acid composition values were converted into compositional similarity scores using principal component analysis, retaining the maximum number of functions for which the chi-square test of Wilks' lambda was significant at $P < 0.05$. We analysed these physico-chemical variables of the pitviper venom $PLA_2$s by DFA in SPSS v.14, using functional activities as groups and individual $PLA_2$ toxins as cases. Data on functional activity were primarily gathered from UniProt. However, it has previously been noted that many database protein entries are not annotated with function (Tan et al., 2003), there are no actively maintained databases specifically for snake venom toxins, and the only current database on animal toxins has limited functionality (Jungo et al., 2012). Therefore, we also carried out searches of the primary literature using GoPubMed (www.gopubmed.org). Reported functional activities of $PLA_2$s are very varied; 15 are listed by Kini (1997) while Doley et al. (2009), mention at least 12 distinct activities. For simplicity, we reduced the number of activities to the six most commonly reported, i.e., neurotoxic, myotoxic, antiplatelet, anticoagulant, oedematous, and hypotensive. Variables were entered together and posterior probabilities of group membership (including for the ungrouped proteins, which did not take part in the discrimination, but whose position relative to the calculated axes was also calculated) were saved.

### 2.5. Profile-based methods

A sequence profile represents the information contained in a multiple sequence alignment as a table of position-specific symbol comparison values and gap penalties. The profile-based neighbour-joining (PNJ) method is a means of obtaining more resolution in a large tree by successively collapsing clusters supported above a certain user-determined value into a summary profile. It is claimed to be as accurate as Bayesian methods, but much more computationally efficient (Müller et al., 2004). We used ProfDistS v0.9.8 (Wolf et al., 2008), with general time-reversible distances based on the VTML model, which models protein evolution as a Markov process (Müller and Vingron, 2000). Profiles were built for clusters with either sequence identity above 97% or bootstrap values (from 500 bootstraps of the initial NJ tree) of greater than 70% in an iterative process (Merget and Wolf, 2010). The resulting PNJ tree was rooted and annotated in Dendroscope 3 (Huson and Scornavacca, 2012). It is important to note that the resulting tree reflects the degree of structural similarity among amino-acid sequences, and will not necessarily reflect evolutionary relationships among the sequences (i.e., it is not a gene tree) since the non-coding parts of the gene may be quite divergent.

### 2.6. Comparison with other prediction methods

A multitude of computational tools are available for the prediction of molecular function based on *de novo* protein sequences (Punta and Ofran, 2008). The more powerful programs combine several approaches. One of these, Protfun (available at http://www.cbs.dtu.dk/services/ProtFun-2.2/), integrates 14 different sequence-based prediction methods, using attributes such as number of negative and positive residues and predicted transmembrane helices, into final predictions of the cellular role, enzyme class (if any), and selected Gene Ontology (GO) categories of the submitted sequence (Juhl-Jensen et al., 2003). As there is a limit of 50 sequences on the server, we assembled a file containing 49 sequences of proteins, in which experimentally determined functions matched the predictions of the DFA ($PP > 0.8$), plus four additional protein sequences with no experimentally determined function, but which the DFA predicted to have a hypotensive or oedematous function with $PP > 0.9$. We also used another multiple-approach protein function prediction engine, EFICAz2.5 available at http://cssb.biology.gatech.edu/skolnick/webservice/EFICAz2/index.html. This combines predictions from six different methods developed and optimised to achieve high prediction accuracy (Narendra and Skolnick, 2012). However, the server takes only one sequence at a time, which limits its utility for large-scale protein discovery projects. Finally, we tested a method employing a similar approach to ours in that it uses features derived from primary sequence such as such as normalised Van der Waals volume, polarity, charge and surface tension. However, rather than employing these measures directly, they are converted into three descriptors which reflect the global composition of each of these properties, and these descriptors are then combined into a feature vector, achieving accuracy in the range 69.1–99.6% (Cai et al., 2003). For the enzyme class to which the $PLA_2$s belong (EC3.1), a sensitivity of 71.1% and specificity of 90.6% is claimed. The server is available at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

To our knowledge, only a handful of other studies have attempted to develop bioinformatic tools specifically for prediction of the biological properties of snake venom $PLA_2$ proteins. Two of these focused on neurotoxins only (Saha

and Raghava, 2007; Siew et al., 2004), one on distinguishing between myotoxins and neurotoxins (Pazzini et al., 2005), and another (Chioato and Ward, 2003) was applied to myotoxins, neurotoxins and anticoagulants. Although these were mostly accompanied by publicly-available programs, only one of these is currently accessible. Consequently, we could only test the predictive power of NTXpred (Saha and Raghava, 2007) available at www.imtech.res.in/raghava/ntxpred/. According to the authors, this server allows users to predict neurotoxins from non-toxins with 97.72% accuracy, allows the classification of neurotoxic proteins by their organismal source with 92.10% overall accuracy, and by function (e.g., ion channel blockers, acetylcholine receptor blockers etc.) with 95.11% overall accuracy. Furthermore, it claims that users can sub-classify ion-channel inhibitors by type with 75% overall accuracy. The interface is simple and limited to the input of one sequence at a time. Consequently, we selected two to three toxins of each functional type to test, from the file described above.

## 3. Results

### 3.1. Description of novel $PLA_2$ proteins

Full-length sequences of $PLA_2$ genes ranging in length between 1832 and 2001 bp were obtained from 24 individuals of 20 nominal species. The minimum difference required for acceptance of variants as non-PCR artefacts was set at 4 bp. After several putative artefactual recombinants were eliminated from the dataset, it consisted of 94 gene sequences. Putative proteins inferred from the coding regions bore hallmarks of expressed genes, including the presence of a TATA-like box and several putative regulatory elements (Gubenšek and Kordiš, 1997) immediately preceding it at the 5′ end, and the polyA tail at the 3′ end. Several genes detected encoded previously described toxins from protein or cDNA studies. For example, B464_LT6 (UniProtKB: *tbc*) from *Protobothrops* (previously *Zhaoermia*) *mangshanensis* encodes a protein with 99% similarity to zhaoermiatoxin (Mebs et al., 2006), while A54_LT6 from *Calloselasma rhodostoma* (UniProtKB: *tbc*), differs by only a single amino acid near the C-terminus from CRV-W6D49 (Tsai et al., 2000). Several distinct genes (as defined above) recovered from the same individual (e.g., B33) or individuals from different populations of the same species (e.g., two *Cryptelytrops* specimens B117 and B5, from South Vietnam and West Java respectively) were found to encode identical proteins. Additionally, several genes encoded toxins with inferred molecular weights that matched the MW of proteins detected by MS analysis of the crude venom obtained from the same, or related, individual. Finally, 10 genes appeared to encode pseudogenes (with either unusually short or long inferred protein sequences according to the position of the first TAA or TAG codon). Accession numbers, origins, inferred MW and pI, sequence features and matches found for the novel sequences in venom MS profiles are shown in Table S1 of the Supplementary Information.

Putatively translated proteins ($n = 73$) varied from 119 to 124 amino acids, within the range of previously described Group II $PLA_2$s (Kini, 1997) and (with the exception of five proteins which had six disulphide bridges), had the usual seven disulphide bridges. The inferred proteins fell into a number of classes previously described, based on the residue present at the 48th position in the amino-acid sequence. Somewhat confusingly, this position is designated 49 in the numbering system proposed by Renetseder et al. (1985) based on a comparison with bovine pancreatic $PLA_2$, in which residue 15 has been deleted in all sv$PLA_2$s. The commonest variant was D49 ($n = 57$), in which the catalytic site is preserved, with a minority ($n = 6$) being K49 (phospholipase homologues). There were also a number of variants at this position (N:6, H:1, R:2, T:1) that have only rarely been previously reported (Chijiwa et al., 2006; Tsai et al., 2003; Wei et al., 2006). The residue at the sixth position of the amino-acid sequence is also significant in determining some important properties of the protein (Mebs et al., 2006). The largest proportion of sequences fell into the E6 category ($n = 49$, mostly of the D49 type, but also including N, K, R and H49 proteins). Most of the E6 proteins are acidic ($4 > pI > 5.5$), but a few are neutral or weakly basic ($pI = 6.4–8.95$), although all are within the range previously reported for E6 proteins. For additional variants at the 6th position (A, G, R, T, W), see Table S1.

### 3.2. Proteomic study of $PLA_2$ expression

Oxidation products (clearly distinguishable as double peaks differing by 16 Da) were frequently present. Among the 10 samples that had been fractionated, isolated isoforms were found to be up to 20% oxidised. These often formed minor peaks in the LC–ES–MS and were generally absent in the MALDI–TOF spectra. From the 132 venoms examined, at least 83 masses representing putative unique $PLA_2$ isoforms were identified between 13,193 and 14,916 Da. Between two (*Popeia sabahi*, A202, *Ovophis makazayazaya*, A87) and 10 (*Viridovipera gumprechti*, B475) isoforms were found in the 24 samples with both LC–ES and MALDI–TOF–MS data. Between 25 and 100% (mean 70.45%) of isoforms in individual venoms were detected using both methods. Most of the masses which did not occur in both types of spectra were present as minor peaks in LC–ES–MS. About 70% of isoforms detected were scored as a major or minor peak consistently in both analyses. There was no significant difference between repeat spectra of the same venom sample, or from venom samples taken at different times from the same individual, although the relative intensity of different peaks and presence of absence of minor peaks were not consistent in some cases. Out of the 73 proteins inferred from the genomic sequences obtained in this study, 62 (c. 85%) had a putative match in the expressed venom (Table S1). However, several isoforms with different amino-acid sequences have inferred masses that are within 2 Da of each other, which are difficult to discriminate using proteomic methods (Table S1), even the more accurate LC–ES–MS. Only 23 (32%) inferred $PLA_2$ proteins were matched to masses in the venom profile of the same individual from which the genome sequence had been obtained, suggesting that selective expression may account for a large proportion of among-individual

variation in venom profiles. However, it also indicates incomplete sampling of the $PLA_2$ gene content of the genomes investigated.

### 3.3. Biological activities

The application of saline-loaded discs of filter paper caused no haemorrhage and no obvious disturbance to the chick embryos. Discs loaded with *B. jararaca* venom exhibited concentration-dependant haemorrhage, with a threshold concentration of 1.0 μg in 2.0 μl. The area of haemorrhagic corona increased with venom concentration and was maximal at a concentration of 3 μg in 2.0 μl, while the time taken for the corona to form fell. From these data, a ranking of haemorrhagic potential was calculated (Table 1). Of the venom samples tested, B135, T221, T61, T61 (fraction 16), T61 (fraction 20) had no detectable haemorrhagic activity. Two samples (B475 and B22) were highly active, as active as the standard haemorrhagic venom (*B. jararaca*). Venom samples from B208, B33, B67 and B5 were moderately active (compared to *B. jararaca*), while those from B8, B469 and A229 were of low haemorrhagic activity. Myotoxic activity was rare and usually mild. Only T224, T221 and T61 (fraction 20) were clearly myotoxic although B526 and T208 were mildly myotoxic. Oedema was common, but non-specific (Table 1). Clear evidence of neurotoxicity was seen only with T61 (fraction 20) (Table 1, Fig. S1).

### 3.4. Prediction of function: discriminant analysis

The total dataset contained 253 non-redundant protein sequences (Fig. 1). The alignment is available in the Dryad data depository (doi:10.5061/dryad.16pg7). The first four factors describing amino acid composition were retained. These principal components, referred to as PC1-4(comp) hereafter, summarised 16.7, 14.0, 10.1, and 9.5% of variation respectively. Ninety-five proteins had known functions that could be assigned to one of six major functions. However, anticoagulant and antiplatelet functions were subsequently combined into a "haemotoxic" category after preliminary analyses showed that no physico-chemical property or PC(comp) could distinguish between these groups (Tamhane's post-hoc test). Final sample sizes were: Myotoxic: 30; Haemotoxic: 19; Neurotoxic: 26; Hypotensive: 7; Oedematous: 15. Neurotoxic $PLA_2$s frequently also show myotoxicity (Montecucco et al., 2008), but were classed as neurotoxic rather than myotoxic for the purpose of this analysis. Robust tests (Brown-Forsythe) for the equality of means showed all variables apart from PC2(comp) showed significant differences among groups. The four resulting discriminant functions (Table 2) contained 69.1, 13.5, 11.1, and 6.3% of variation respectively. Another 158 proteins which had no known function were plotted on the resulting axes (Fig. 2) and colour-coded by their posterior probabilities of belonging to one of the functional groups (Table S2). All groups, except for haemotoxic and hypotensive proteins, were successfully discriminated on two axes (Fig. 2A). DF1 largely reflects the difference in pI, with haemotoxic/hypotensive proteins being acidic, myotoxic, neurotoxic and most oedematous proteins being basic. However, notably some oedematous proteins can be distinguished from myotoxic ones by being more neutrally charged at pH7. DF2 (Table 2, Fig. 2A) largely distinguishes a smaller group of oedematous proteins on the basis of PC3(comp), with oedematous toxins having lower amounts of phenylalanine, arginine and tyrosine, and higher amounts of methionine and valine. DF3 (not shown)

**Table 1**
Biological activities of several characterised crude venoms and three isolated isoforms.

| Venom ID | Species | Time to corona | Rank | Size of corona ($mm^2$) | Rank | Rank overall (haemorrhagic activity) | Oedema | Myotoxicity | AchR | $S^1$ physin | % innervated |
|---|---|---|---|---|---|---|---|---|---|---|---|
| +ve control | *Bothrops jararaca* | 26.5 | 5 | 17 | 2 | 2= | n/a | n/a | n/a | n/a | n/a |
| −ve control | – | n/a | n/a | n/a | n/a | n/a | Slight | X | 43 | 40 | 93 |
| Notexin | *Notechis scutatus* | n/a | n/a | n/a | n/a | n/a | √ | √ | n/a | n/a | n/a |
| β Bungarotoxin | *Bungarus* spp. | n/a | n/a | n/a | n/a | n/a | n/a | n/a | 62 | 16 | 26 |
| A229 | *Cryptelytrops albolabris* | 66 | 10 | 5.5 | 9 | 10 | X | X | 28 | 28 | 100 |
| T208 | *Viridovipera stejnegeri* | 24 | 4 | 8 | 7 | 4= | √ | √ | 28 | 28 | 100 |
| B475 | *V. gumprechti* | 20 | 2 | 19 | 1 | 1 | √ | X | 72 | 66 | 92 |
| B8 | *C. insularis* | 44 | 9 | 16 | 3 | 8 | X | X | 52 | 48 | 92 |
| B22 | *C. albolabris* | 17 | 1 | 8.6 | 6 | 2= | X | X | 15 | 12 | 80 |
| B526 | *"Ovophis" okinavensis* | 50 | 11 | 3.2 | 11 | 11 | X | Slight | 20 | 18 | 90 |
| B5 | *C. albolabris* | 30 | 7 | 11 | 4 | 4= | X | X | 33 | 30 | 91 |
| B33 | *Parias hageni* | 27 | 6 | 8.8 | 5 | 4= | X | X | 42 | 38 | 90 |
| B67 | *C. cardamomensis* | 21 | 3 | 7.7 | 8 | 4= | X | X | – | – | – |
| B469 | *Popeia sabahi* | 43 | 8 | 5.3 | 10 | 9 | X | X | 42 | 40 | 95 |
| B135 | *Tropidolaemus wagleri* | Not haemorrhagic | | | 12= | 12= | X | X | 75 | 62 | 83 |
| T221 (crude) | *V. stejnegeri* | Not haemorrhagic | | | 12= | 12= | √ | √ | 54 | 48 | 89 |
| T221 (Q6H3D7)) | *V. stejnegeri* | Not haemorrhagic | | | 12= | 12= | – | – | – | – | – |
| T61 (crude) | | Not haemorrhagic | | | 12= | 12= | – | – | 22 | 20 | 91 |
| T61/16 (Q6H3D4) | | Not haemorrhagic | | | 12= | 12= | X | X | 13 | 13 | 100 |
| T61/20 (D31778) | | Not haemorrhagic | | | 12= | 12= | √ | √ | 20 | 0 | 0 |

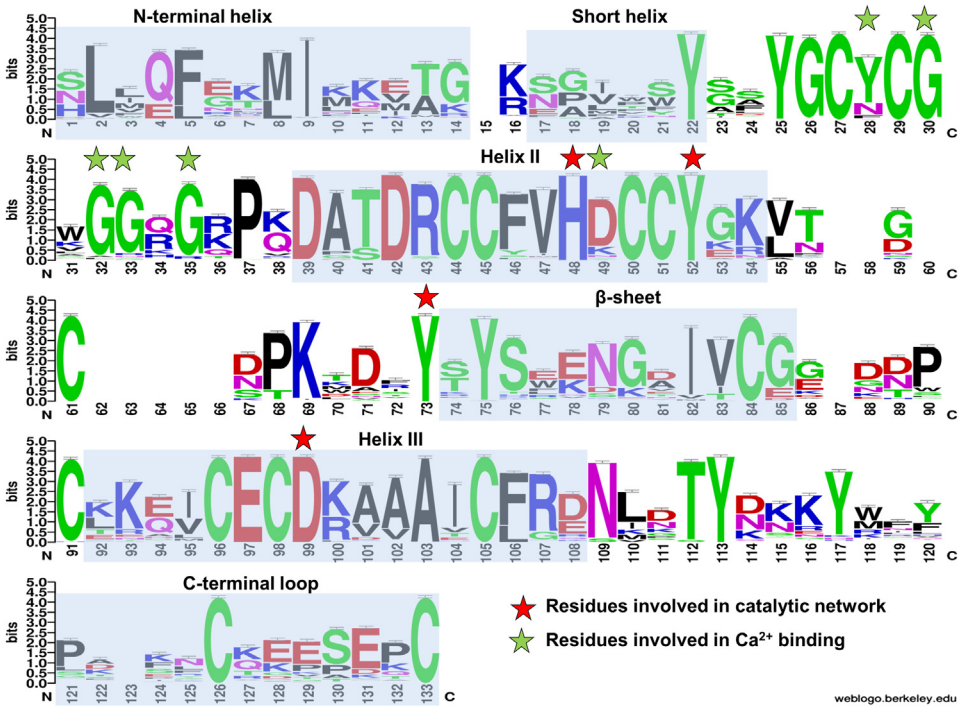


**Fig. 1.** Sequence logo constructed from the 251 $PLA_2$ sequences included in this study, showing conserved and variable regions. Standard amino-acid single-letter codes are used, and the size of the letter indicates the degree of conservation of the sequence at that position. In order to preserve the standard $PLA_2$ numbering system (Renetseder et al., 1985), gaps have been introduced even though no $svPLA_2$ has a residue present at that position, and a few extra inserted residues found in a minority of $svPLA_2$s (fewer than two) have been deleted. Major three-dimensional structural features are shown as blue boxes and residues participating in the catalytic activity of the enzyme have been indicated using stars.

is influenced by a contrast between net charge and pI, and further distinguishes myotoxins proteins from oedematous proteins and neurotoxins, with myotoxins displaying a lower net charge for a given pI than the other types. Finally, hypotensive $PLA_2$s are distinguished from haemotoxins on the fourth axis (Fig. 2B), by a lower pI, a higher proportion of leucine and lycine and a lower amount of alanine, cysteine, glutamic acod and glutamine, being less thermostable and more hydrophilic. Of original grouped toxins, 72.6% were correctly classified while cross-validation correctly classified 60% of toxins. Of the 27 known myotoxic proteins, 21 (78%) were correctly predicted. The prediction accuracy of known hypotensive proteins is 86% (6 out of 7), while neurotoxic and oedematous proteins were both correctly predicted in 62% of cases. Haemotoxic proteins were correctly predicted in 74% of cases.

**Table 2**
Standardised eigenvector coefficients for canonical discriminant functions on viperid $PLA_2$ composition and physico-chemical properties. The coefficients shown in bold contribute the most to the respective discriminant function (DF).

| Function variable | DF1 | DF2 | DF3 | DF4 |
|---|---|---|---|---|
| GRAVY | −0.345 | −0.504 | −0.542 | **−0.784** |
| Instability index | 0.095 | 0.047 | 0.320 | 0.409 |
| Aliphatic index | 0.265 | **0.743** | 0.205 | **0.882** |
| Amp score | 0.164 | 0.018 | 0.655 | −0.188 |
| Net charge | **−0.851** | 0.598 | **2.184** | **−0.742** |
| NO. A.A. | −0.010 | 0.008 | 0.467 | 0.591 |
| pI | **1.789** | −0.643 | **−1.634** | **0.920** |
| MW | −0.042 | 0.169 | −0.784 | −0.166 |
| PC1(comp) | −0.053 | 0.497 | −0.235 | 0.661 |
| PC2(comp) | 0.068 | 0.506 | −0.295 | **0.908** |
| PC3(comp) | −0.122 | **0.974** | 0.087 | −0.281 |
| PC4(comp) | −0.340 | −0.136 | −0.140 | 0.494 |

### 3.5. *Profile-based clustering*

The profile neighbour-joining tree (Fig. 3) shows good correspondence between cluster membership and known and/or predicted functions, although much of the deeper structure of the tree is not supported by bootstrap analysis. For example, only one known myotoxin lies outside a cluster containing proteins with similar functions. A fundamental split between proteins with a mainly haemotoxic (and hypotensive) function and proteins having oedematous, myotoxic or neurotoxic activity is evident. Apart from the distinct clustering of viperine sequences (clusters A and B) there is no particularly strong signal of taxonomy in the tree (e.g., cluster D, which largely groups toxins from rattlesnakes, also contains toxins from the Old World genera *Ovophis* and *Gloydius*). Interestingly, hypotensive $PLA_2$s seem to be structurally similar in viperines,

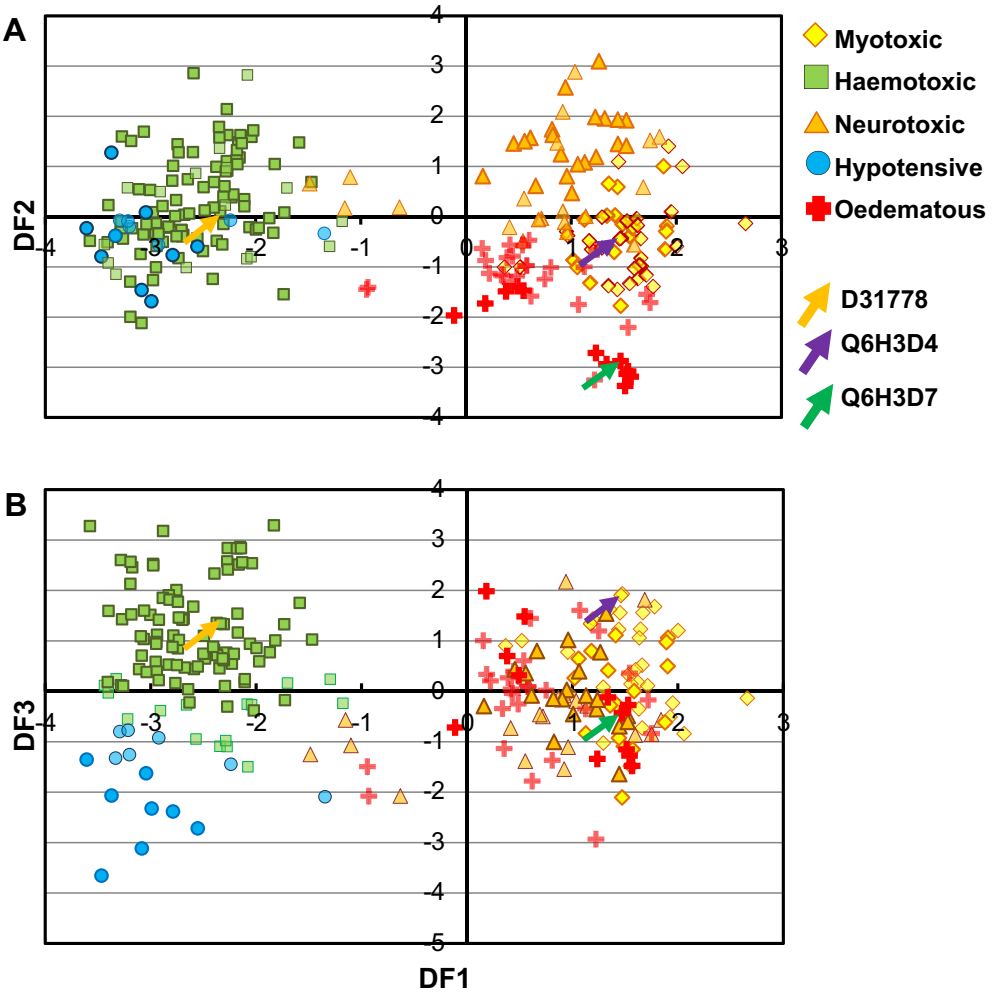


**Fig. 2.** Discriminant function plot showing A) DF1 against DF2 and B) DF1 against DF4. All functional groups of $PLA_2$ isoforms, except "haemotoxic" and "hypotensive", are successfully discriminated on the first two axes while the remaining groups are successfully discriminated on DF4. The intensity of the colour of the symbol distinguishes proteins with a posterior probability (PP) > 80% of belonging to that group (darker), from those with 50% < PP>80% (paler). Arrows indicate the position of three isoforms from the venom of *Viridovipera stejnegeri*, whose bioactivity was investigated in this study.

occurring in only cluster A, despite disparate specific origins. However, in the crotalines, they appear independently among different clusters, and are always very similar to a haemotoxic protein. Similarly, oedematous activity and myotoxicity are also closely related, with whole clusters being identified containing proteins known/predicted to have one of these activities (e.g., clusters C and E, Fig. 3). The independent evolution of myotoxins is indicated by their occurrence in each of the two clusters of viperine $PLA_2$s (A and B) and in several distinct clusters of crotaline toxins (C, D, E and predicted, but not confirmed, in some other clusters as well). Although not well illustrated in the figure, which shows only one function for each toxin, many neurotoxins from pitvipers can also display myotoxicity. This is true of many of the known neurotoxins in cluster C and D, which may explain many of the discrepancies observed between known and predicted function in these clusters. A large number of the inferred haemotoxins examined, however, are not strongly structurally related and fall into a number of small clusters whose relationships are unclear. Within these are located the small clusters of $PLA_2$s with known hypotensive activity and, perhaps more surprisingly, two known neurotoxic $PLA_2$s. These are not predicted as neurotoxins by DFA, and may have acquired neurotoxicity recently and independently.

### 3.6. Comparison with other prediction methods

Results from Protfun 2.2 did not correspond with expected classifications. Out of the 49 sequences submitted, almost half (20) were predicted to be non-enzymes. Of these, 70% are indeed likely to be $PLA_2$ homologues due to substitutions present at the critical 49th residue. Overall the accuracy of predicting enzyme activity was 85.7%, but none were correctly classed as Hydrolases (EC 3.-.-.-); instead, six were predicted to be isomerases, and no predictions were provided for the remainder. EFICAz[2.5], on the other hand, correctly classified all the sequences tested as phospholipase $A_2$ enzymes (EC 3.1.1.4) with high confidence, but each protein sequence took nearly two hours to

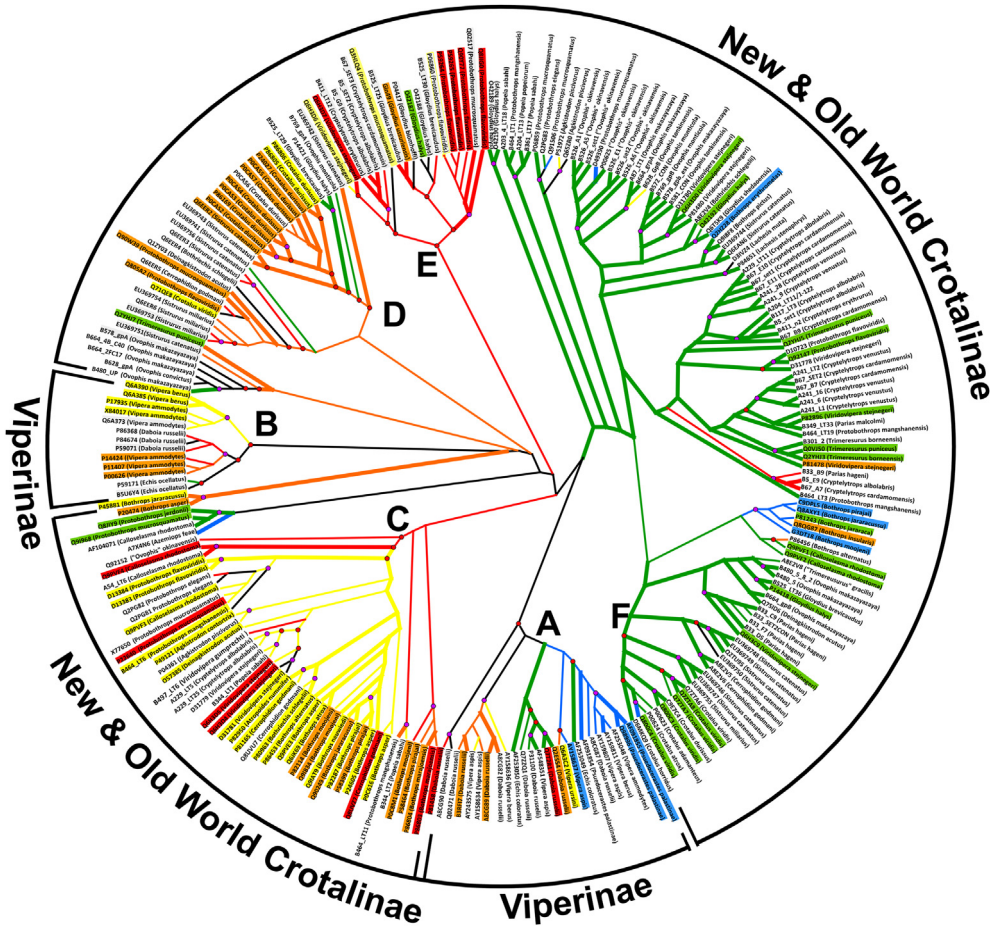


**Fig. 3.** A profile neighbour-joining (PNJ) tree of all Group II svPLA$_2$ protein sequences illustrated as a radial cladogram. Branches have been colour coded according to the functional predictions derived from the DA, with colours corresponding to those used in Fig. 2. Thick lines represent proteins with a PP >80% of having that function, while thin lines represent 50% < PP>80%. Black lines indicate proteins which had no clear assignment (PP>50%), or in the case of interior branches, where descendant sequences had multiple functions. Nodes marked by red circles represent those that are supported by bootstrap values > 50% in the final tree, while those marked by a purple circle represent sequences which have been united into a super-profile by several iterations of PNJ. Peripheral labels indicate the unique identifier of each protein and the species from which it was identified. Labels are coloured according to known (experimentally determined) functions. The outer circle indicates the position of two clades of Viperinae (true viper) sequences, which are clearly distinguished from all other PLA$_2$s of crotaline (pit viper) origin. Labels A–F identify clusters that are supported by bootstrap values >50% and which are discussed in the text.

be processed. SVMProt also returned a prediction of EC 3.1.-.- (Hydrolases – Acting on Ester Bonds) with 95.9% accuracy. For a further two proteins, the classification with the highest probability was "all lipid-binding proteins". However, as pointed out earlier, information on enzyme activity is of limited utility when dealing with multifunctional proteins such as the svPLA$_2$s. NTXpred tools varied in their prediction of source, function and specificity (Table S4) but all PLA$_2$s tested were predicted to be neurotoxins. In order to investigate the prediction accuracy further, the amino acid sequence was randomly mutated and the prediction tools run after each mutation. At least two out of the 14 Cys residues that form the crucial backbone of the protein had to be mutated before the amino-acid + length tool predicted a non-toxin, at least four Cys residues had to be mutated before the dipeptide-based tools failed to predict a neurotoxin, and all Cys could be mutated and still obtain a neurotoxin prediction from the "amino-acid sequence only" tool. If these cysteine residues were untouched, the entire remaining amino-acid sequence could be randomly changed without changing the prediction.

## 4. Discussion

The prediction of function from protein sequence in the toxic PLA$_2$s is especially challenging, yielding few insights despite decades of work in this field. To some extent, this lack of progress can be attributed to incomplete analysis and lack of standardisation in the toxinological literature. For example, while reported activities of phospholipases are very varied (Doley et al., 2009), few have been extensively studied and individual toxins are rarely tested for all possible activities. Thus, it cannot be ascertained whether the toxin also shows activities additional to the experimentally demonstrated ones, which may account for some apparent misclassifications in predictive methods such as

those investigated here. Additionally, assay methods vary considerably and some are far more sensitive than others. For example, measuring the resting membrane potential in the mouse phrenic nerve-diaphragm preparation was found to be around 100-fold more sensitive than the commonly-used creatine kinase release assay for studying myotoxicity (Aragão et al., 2009). In addition, the same pharmacological effect can be induced through different pathways (Miyabara et al., 2006; Moreira et al., 2008; Zhou et al., 2008).

Although several attempts have been made to provide a predictive framework for $PLA_2$ toxins, none of these have gained widespread use and published databases and servers eventually become unavailable and/or outdated. As we have illustrated, a number of more general methods (not designed specifically for toxins) lack predictive power, while specific tests to identify toxins (Saha and Raghava, 2007) fail to distinguish between different toxic functions. Among the methods not currently accessible, some reported success in prediction of myotoxic, presynaptic neurotoxic and anticoagulant functions was achieved by examining subsets of highly similar toxins (found by sequence similarity searches of databases) (Chioato and Ward, 2003). However, the assumption that sequences with high similarity share a similar function has been shown to be flawed in this study, where we find that similar functions may have evolved independently in structurally different sequences, while some novel functions have arisen among clusters of highly similar sequence, making it difficult to identify functional relationships among sequences grouped by similarity alone. This is illustrated by clusters C and D in Figs. 3 and 4, both containing largely myotoxic/oedematous $PLA_2$s as well as a number of neurotoxic $PLA_2$s. However, this underlying similarity in physiological effect is clearly achieved through different biochemical pathways, as $PLA_2$s in cluster D are all highly catalytically active, and the neurotoxicity is achieved through dimerisation with a non-toxic chaperone protein. Members of cluster C, on the other hand, all have mutations that have abolished or considerably reduced the catalytic activity, and when neurotoxic, can express this activity in the monomeric form. The presence of both these activities in both these structurally distinct clusters may be one reason that considerable overlap was found in the surface residues implicated in myotoxicity and neurotoxicity (Chioato and Ward, 2003). The paucity of existing data on some particular functions (e.g., hypotensive $PLA_2$s, where we were only able to find experimental evidence for this activity for seven isoforms among all viperids) also challenges the ability of any method to classify them.

A particularly encouraging feature of the current analysis is the good agreement between cluster membership in the

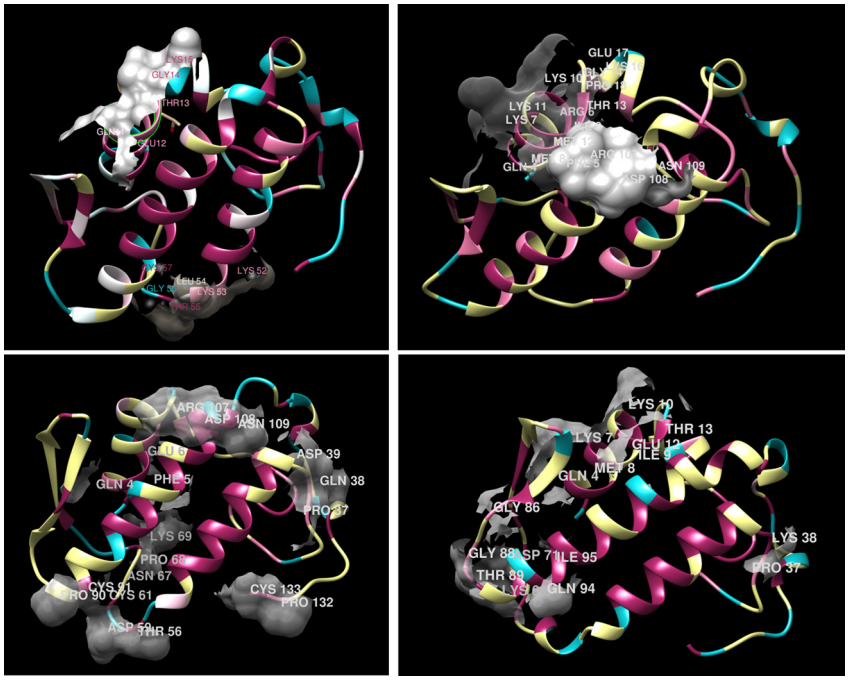


**Fig. 4.** Molecular models showing conserved regions (excluding catalytic and calcium binding sites) in crotaline svPLA$_2$ clades C, D, E and F (as defined in Fig. 3). Top left: clade C (mostly myoxins but with some predicted oedematous and some known neurotoxins); top right: clade E (mostly oedematous); Bottom left: clade F (mostly anticoagulant), bottom right: clade D (mostly neurotoxic but also showing myotoxic activity, non-toxic without chaperone molecule). The molecules are all oriented with beta-sheet on the extreme left (front), the n-terminal helix towards the left (behind), and the c-terminal loop on the extreme right, for ease of comparison. Surfaces and residue labels are shown for highly conserved and exposed residues only. Conservation scores were calculated using the ConSurf server (Ashkenazy et al., 2010; Landau et al., 2005) using the reported PDB structures of 3I03 *(Bothrops jararacussu)*, 1VIP (*Daboia russelii*), 2QOG (*Crotalus durissus*), and 1PP2 (*Crotalus atrox*) respectively. The darkest maroon colours are the most conserved while the darkest blue colours are the most variable. Yellow indicates that the sample size was too small for a confident prediction to be made at that site. The images were produced using the USCF Chimera package (Pettersen et al., 2004). It suggests that each clade has largely non-overlapping active sites.

PNJ trees, based on sequence profiles, and the functional predictions from the DFA based on physico-chemical properties, which have different underlying bases. We also found good internal consistency between our predictions and *in vitro* tests of activity. For example, venom from specimen T208 (*V. stejnegeri* from Taiwan) is known from the proteomic analysis to contain major $PLA_2$s that match the MW of sequenced isoforms A241_9 and B344_LT2. The third major isoform present matches the MW of Q6H3D4, which was tested as part of this study and showed no distinct activity. This matches the prediction from the DFA which could not clearly assign it to any functional cluster. The crude venom showed haemorrhagic, oedematous and myotoxic activity. A241_9 is predicted with a PP of 0.99 to be a haemotoxin while B344_LT2 is predicted (PP = 0.66) to be a myotoxin, thus the demonstrated activity of the whole venom is entirely consistent with the predictions of the functional activity of its main constituent $PLA_2$s. Similarly, the activity of the crude venom from B469, B475, B526, B5, B33 and B67 is entirely consistent with the predicted activity of at least some of the major $PLA_2$ toxins that they contain. The activity of the venom from B8 (*Cryptelytrops insularis*) is partly consistent, in that it is known to contain isoforms that have predicted activities that are not shown by the whole venom. However, in this case, the only major toxin (matching B5_set2 in MW) is predicted to be haemotoxic (PP = 0.94), which matches the activity of the crude venom, while the isoform matching A229_LT5 (with predicted myotoxic activity) is only a minor constituent of the venom (data from the LC–ES–MS). A more inexplicable inconsistency between predicted and demonstrated functions is found in the case of the crude venom of A229 (*Cryptelytrops albolabris*), which showed only slight haemorrhagic activity and no other activity. From the LC–ES–MS profile, we know that this venom contains seven major isoforms of $PLA_2$, six of which have been identified in this study (these are A229_LT5, A229_LT11, A241_28, B464_LT11, B480_UP, and B769_gpB), and another which remains unidentified. Of these, A229_LT11, A241_28 and B769_gpB have predicted haemotoxic activity (PP > 0.9), but B464_LT11 has predicted neurotoxic activity (PP = 0.82) and A229_LT5 has predicted myotoxic activity (PP = 0.6). There may be synergistic effects among this complex cocktail of similar toxins that masks some of these activities in the crude venom. This may also be the reason for a dramatic inconsistency between the results of the functional assays on whole venom and the isolated toxins in the case of D31778, which was isolated from the venom of T221 (*V. stejnegeri*). The isolated toxin shows very high neurotoxic activity which exceeded that of the positive control used, yet the whole venom shows no such activity. In this case, the neurotoxicity of D31778 also fails to be predicted by the DFA (which in fact predicts it to be a haemotoxin with very high probability), and in the PNJ tree, is clustered among other isoforms similarly predicted to be haemotoxins. It is therefore extremely interesting that another isoform from *V. stejnegeri* (P81478) has been independently demonstrated to be neurotoxic (Fukagawa et al., 1993), yet also fails to be predicted as such by the current methods. This suggests that the another use of the approach outlined here may be to highlight discrepancies between expected and actual functions which, by departing from the norm, may yield unique information about the gain and loss of major functions in these versatile proteins.

In conclusion, our study highlights the importance of considering biological meaningful features of proteins for detailed understanding of their biological activities. With the number of venomous animals running into many tens of thousands, the search for bioactive compounds as leads in the pharmaceutical industry in these venoms will need to be organised for maximum efficiency. A method of providing an initial hypothesis of function of a novel product that is capable of highlighting the independent acquisition of similar functions by toxins of different sequence, that may act through different pathways, could be a valuable tool in choosing such lead compounds for further investigation.

## Acknowledgements

We thank Wendy Grail, Carlotta Ercolani (Bangor), and Tracey Davey (Newcastle), for their assistance in the laboratory, and Karen Dawson for making available the unpublished $PLA_2$ gene sequences from *Ovophis* species from her PhD thesis.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.toxicon.2013.06.019.

## Conflict of interest statement

The authors declare that there are no conflicts of interest. This work was supported by the Natural Environment Research Council of the United Kingdom (grant number NER/A/S/2001/01217 to AM, RST and JBH). The sponsor had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

## References

Aragão, E.A., Randazzo-Moura, P., Rostelato-Ferreira, S., Rodrigues-Simioni, L., Ward, R.J., 2009. Shared structural determinants for the calcium-independent liposome membrane permeabilization and sarcolemma depolarization in Bothropstoxin-I, a LYS49-$PLA_2$ from the venom of *Bothrops jararacussu*. Int. J. Biochem. Cell. Biol. 41, 2588–2593.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N., 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 38, W529–W533. http://dx.doi.org/10.1093/nar/gkq399.

Bazaa, A., Pasquier, E., Defilles, C., Limam, I., Kessentini-Zouari, R., Kallech-Ziri, O., Battari, A.E., Braguer, D., Ayeb, M.E., Marrakchi, N., Luis, J., Tailleux, L., 2010. MVL-PLA2, a snake venom phospholipase $A_2$, inhibits angiogenesis through an increase in microtubule dynamics and disorganization of focal adhesions. PLoS One 5 (4), e10124. http://dx.doi.org/10.1371/journal.pone.0010124.

Boldrini-França, J., Rodrigues, R.S., Fonseca, F.P., Menaldo, D.L., Ferreira, F.B., Henrique-Silva, F., Soares, A.M., Hamaguchi, A., Rodrigues, V.M., Otaviano, A.R., Homsi-Brandeburgo, M.I., 2009. *Crotalus durissus collilineatus* venom gland transcriptome, analysis of gene expression profile. Biochimie 5, 586–595.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z., 2003. SVM-Prot, web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 31, 3692–3697.

Calvete, J.J., Sanz, L., Pérez, A., Borges, A., Vargas, A.M., Lomonte, B., Angulo, Y., Gutiérrez, J.M., Chalkidis, H.M., Mourão, R.H., Furtado, M.F.,

Moura-Da-Silva, A.M., 2011. Snake population venomics and antivenomics of *Bothrops atrox*: Paedomorphism along its transamazonian dispersal and implications of geographic venom variability on snakebite management. J. Proteomics 74, 510–527.
Chatrath, S.T., Chapeaurouge, A., Lin, Q., Lim, T.K., Dunstan, N., Mirtschin, P., Kumar, P.P., Kini, R.M., 2011. Identification of novel proteins from the venom of a cryptic snake *Drysdalia coronoides* by a combined transcriptomics and proteomics approach. J. Proteome Res. 10, 739–750.
Chijiwa, T., Tokunaga, E., Ikeda, R., Terada, K., Ogawa, T., Oda-Ueda, N., Hattori, S., Nozaki, M., Ohno, M., 2006. Discovery of novel (Arg49) phospholipase $A_2$ isozymes from *Protobothrops elegans* venom and regional evolution of Crotalinae snake venom phospholipase $A_2$ isozymes in the southwestern islands of Japan and Taiwan. Toxicon 48, 672–682.
Chioato, L., Ward, R.J., 2003. Mapping structural determinants of biological activities in snake venom phospholipases $A_2$ by sequence analysis and site directed mutagenesis. Toxicon 42, 869–883.
Clark, W.T., Radivojac, P., 2010. Analysis of protein function and its prediction from amino acid sequence. Proteins 79, 2086–2096.
Coulthard, L.G., Costello, J., Robinson, B., Shiels, I.A., Taylor, S.M., Woodruff, T.M., 2011. Comparative efficacy of a secretory phospholipase $A_2$ inhibitor with conventional anti-inflammatory agents in a rat model of antigen-induced arthritis. Arthritis Res. Ther. 13, R42.
Creer, S., Malhotra, A., Stöcklin, R., Favreau, P., Thorpe, R.S., et al., 2003. Genetic and ecological correlates of intraspecific variation in pitviper venom composition detected using Matrix-Assisted Laser Desorption Time-of-Flight Mass Spectrometry, MALDI-TOF-MS, isoelectric focusing, $PLA_2$. J. Mol. Evol. 56, 317–329.
Danse, J.M., Gasparani, S., Menez, A., 1997. Molecular biology of snake venom phospholipases $A_2$. In: Kini, R.M. (Ed.), Venom Phospholipase $A_2$ Enzymes, Structure, Function and Mechanism. Wiley, Chichester, pp. 29–72.
Dawson, K., Thorpe, R.S., Malhotra, A., 2010. Estimating genetic variability in non-model taxa: a general procedure for discriminating sequence errors from actual variation. PLoS One 5 (12), e15204.
Dixon, R.W., Harris, J.B., 1999. Nerve terminal damage by β-bungarotoxin-its clinical significance. Am. J. Pathol. 154, 447–455.
Doley, R., Zhou, X., Kini, R.M., 2009. Snake venom phospholipase $A_2$ enzymes. In: Mackessy, S.P. (Ed.), Handbook of Venoms and Toxins of Reptiles. CRC Press, Florida, pp. 173–205. http://dx.doi.org/10.1201/9781420008661.ch8.
Durban, J., Juárez, P., Angulo, Y., Lomonte, B., Flores-Diaz, M., Alape-Girón, A., Sasa, M., Sanz, L., Gutiérrez, J.M., Dopazo, J., Conesa, A., Calvete, J.J., 2011. Profiling the venom gland transcriptomes of Costa Rican snakes by 454 pyrosequencing. BMC Genomics 12, 259.
Edgar, R.C., 2004. MUSCLE, multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Engelhardt, B.E., Jordan, M.I., Srouji, J.R., Brenner, S.E., 2011. Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res. 21, 1969–1980.
Fenard, D., Lambeau, G., Valentin, E., Lefebvre, J.C., Lazdunski, M., Doglio, A., 1999. Secreted phospholipases $A_2$, a new class of HIV inhibitors that block virus entry into host cells. J. Clin. Invest. 104, 611–618.
Fukagawa, T., Nose, T., Shimohigashi, Y., Ogawa, T., Oda, N., Nakashima, K., Chang, C.C., Ohno, M., 1993. Purification, sequencing and characterization of single amino acid-substituted phospholipase A2 isozymes from *Trimeresurus gramineus*, green habu snake. venom. Toxicon 31, 957–967.
Glenn, T.C., 2011. Field guide to next generation DNA sequencers. Mol. Ecol. Resour.. http://dx.doi.org/10.1111/j.1755-0998.2011.03024.x.
Gowda, C.D., Rajesh, R., Nataraju, A., Dhananjaya, B.L., Raghupathi, A., Gowda, T.V., Sharath, B.K., Vishwanath, B.S., 2006. Strong myotoxic activity of *Trimeresurus malabaricus* venom, role of metalloproteases. Mol. Cell Biochem. 282, 147–155.
Gubenšek, F., Kordiš, D., 1997. Venom phospholipase $A_2$ genes and their molecular evolution. In: Kini, R.M. (Ed.), Venom Phospholipase $A_2$ Enzymes, Structure, Function and Mechanism. Wiley, Chichester, pp. 73–95.
Harris, J.B., Johnson, M.A., Karlsson, E., 1975. Pathological responses of rat skeletal muscle to a single subcutaneous injection of a toxin isolated from the venom of the Australian tiger snake, *Notechis scutatus scutatus*. Clin. Exp. Pharmacol. Physiol. 2, 383–404.
Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol.. http://dx.doi.org/10.1093/sysbio/sys062. Software freely available from: http://www.dendroscope.com.
Jiang, Y., Li, Y., Lee, W., Xu, X., Zhang, Y., Zhao, R., Zhang, Y., Wang, W., 2011. Venom gland transcriptomes of two elapid snakes, *Bungarus multicinctus* and *Naja atra*, evolution of toxin genes. BMC Genomics 12, 1.
Juhl-Jensen, L., Stærfeldt, H.H., Brunak, S., 2003. Prediction of human protein function according to gene ontology categories. Bioinformatics 19, 635–642.
Jungo, F., Bougueleret, L., Xenarios, I., Poux, S., 2012. The UniProtKB/Swiss-Prot Tox-Prot program, a central hub of integrated venom protein data. Toxicon 60, 551–557. http://dx.doi.org/10.1016/j.toxicon.2012.03.010.
Junqueira-de-Azevedo, I.L.M., Ho, P.L., 2002. A survey of gene expression and diversity in the venom glands of the pitviper snake *Bothrops insularis* through the generation of expressed sequence tags, ESTs. Gene 299, 279–291.
Kashima, S., Roberto, P.G., Soares, A.M., Astolfi-Filho, S., Pereira, J.O., Giuliati, S., Faria Jr., M., Xavier, M.A., Fontes, M.R., Giglio, J.R., França, S.C., 2004. Analysis of *Bothrops jararacussu* venomous gland transcriptome focusing on structural and functional aspects. I. Gene expression profile of highly expressed phospholipases $A_2$. Biochimie 86, 211–219.
King, G.F., 2011. Venoms as a platform for human drugs, translating toxins into therapeutics. Expert Opin. Biol. Ther. 11, 1469–1484.
Kini, R.M., 2006. Anticoagulant proteins from snake venoms, structure, function and mechanism. Biochem. J. 397, 377–387.
Kini, R.M. (Ed.), 1997. Venom Phospholipase $A_2$ Enzymes, Structure, Function, and Mechanism. Wiley, Chichester.
Kini, R.M., Iwanaga, S., 1986a. Structure-function relationships of phospholipases. I, prediction of presynaptic neurotoxicity. Toxicon 24, 527–541.
Kini, R.M., Iwanaga, S., 1986b. Structure-function relationships of phospholipases. II, charge density distribution and the myotoxicity of presynaptically neurotoxic phospholipases. Toxicon 24, 895–905.
Koh, D., Armugam, A., Jeyaseelan, K., 2006. Snake venom components and their applications in biomedicine. Cell. Mol. Life Sci. 63, 3030–3041.
Koonin, E.V., 2000. Bridging the gap between sequence and function. Trends Genet. 16, 16.
Kyte, J., Doolittle, R., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.
Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., Ben-Tal, N., 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. 33, W299–W302.
Lomonte, B., Angulo, Y., Moreno, E., 2010. Synthetic peptides derived from the C-terminal region of Lys49 phospholipase $A_2$ homologues from Viperidae snake venoms, biomimetic activities and potential applications. Curr. Pharm. Des. 16, 3224–3230.
Malhotra, A., Thorpe, R.S., 2004. A phylogeny of four mitochondrial gene regions suggests a revised taxonomy for Asian pit vipers, *Trimeresurus* and *Ovophis*. Mol. Phylogenet. Evol. 32, 83–100.
Martin, D.P., Williams, C., Posada, D., 2005. RDP2, recombination detection and analysis from sequence alignments. Bioinformatics 21, 260–262.
Mebs, D., Kuch, U., Coronas, F.I.V., Batista, C.V.F., Gumprecht, A., Possani, L.D., 2006. Biochemical and biological activities of the venom of the Chinese pitviper *Zhaoermia mangshanensis*, with the complete amino acid sequence and phylogenetic analysis of a novel Arg49 phospholipase $A_2$ myotoxin. Toxicon 47, 797–811.
Meenakshisundaram, R., Sweni, S., Thirumalaikolundusubramanian, P., 2009. Hypothesis of snake and insect venoms against human immunodeficiency virus, a review. AIDS Res. Ther. 6, 25.
Merget, B., Wolf, M., 2010. A molecular phylogeny of *Hypnales*, Bryophyta. inferred from ITS2 sequence-structure data. BMC Res. Notes 3, 320.
Miyabara, E.H., Baptista, I.L., Lomonte, B., Selistre-de-Araújo, H.S., Gutiérrez, J.M., Moriscot, A.S., 2006. Effect of calcineurin inhibitors on myotoxic activity of crotoxin and *Bothrops asper* phospholipase $A_2$ myotoxins in vivo and in vitro. Comp. Biochem. Physiol. C 143, 284–294.
Montecucco, C., Gutiérrez, J.M., Lomonte, B., 2008. Cellular pathology induced by snake venom phospholipase $A_2$ myotoxins and neurotoxins, common aspects of their mechanisms of action. Cell. Mol. Life Sci. 65, 2897–2912.
Moreira, V., Gutiérrez, J.M., Soares, A.M., Zamunér, S.R., Purgatto, E., Teixeira, C., de, F., 2008. Secretory phospholipases $A_2$ isolated from *Bothrops asper* and from *Crotalus durissus terrificus* snake venoms induce distinct mechanisms for biosynthesis of prostaglandins E2 and D2 and expression of cyclooxygenases. Toxicon 52, 428–439.
Müller, T., Vingron, M., 2000. Modeling amino acid replacement. J. Comput. Biol. 7, 761–776.
Müller, T., Rahmann, S., Dandekar, T., Wolf, M., 2004. Accurate and robust phylogeny estimation based on profile distances, a study of the Chlorophyceae, Chlorophyta. BMC Evol. Biol. 4, 20.
Narendra, K., Skolnick, J., 2012. EFICAz2.5, application of a high-precision enzyme function predictor to 396 proteomes. Bioinformatics 28, 2687–2688. http://dx.doi.org/10.1093/bioinformatics/bts510.
Ogawa, T., Oda, N., Nakashima, K., Sasaki, H., Hattori, M., Sakaki, Y., Kihara, H., Ohno, M., 1992. Unusually high conservation of

untranslated sequences in cDNAs for *Trimeresurus flavoviridis* phospholipase $A_2$ isozymes. Proc. Natl. Acad. Sci. U. S. A. 89, 8557–8561.
Ohno, M., Ménez, R., Ogawa, T., Danse, J.M., Shimohigashi, Y., Fromen, C., Ducancel, F., Zinn-Justin, S., Le Du, M.H., Boulain, J.C., Tamiya, T., Ménez, A., 1998. Molecular evolution of snake toxins, is the functional diversity of snake toxins associated with a mechanism of accelerated evolution? Prog. Nucleic Acid Res. Mol. Biol. 59, 307–364.
Pazzini, F., Oliveira, F., Guimarães, J.A., Neubauer de Amorim, H.L., 2005. Prediction of myotoxic and neurotoxic activities in phospholipases A2 from primary sequence analysis. In: Setubal, J., Verjovski-Almeida, S. (Eds.), Advances in Bioinformatics and Computational Biology. Springer, Berlin, pp. 194–197.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612.
Prasarnpun, S., Walsh, J., Awad, S.S., Harris, J.B., 2005. Envenoming bites by kraits, the biological basis of treatment resistant neuromuscular paralysis. Brain 128, 2987–2996.
Punta, M., Ofran, Y., 2008. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput. Biol. 4 (10), e1000160. http://dx.doi.org/10.1371/journal.pcbi.1000160.
Renetseder, R., Brunie, S., Dijkstra, B.W., Drenth, J., Sigler, P.B., 1985. A comparison of the crystal structures of phospholipase $A_2$ from bovine pancreas and *Crotalus atrox* venom. J. Biol. Chem. 260, 11627–11634.
Rodrigues, R.S., Boldrini-França, J., Fonseca, F.P., de la Torre, P., Henrique-Silva, F., Sanz, L., Calvete, J.J., Rodrigues, V.M., 2012. Combined snake venomics and venom gland transcriptomic analysis of *Bothropoides pauloensis*. J. Proteomics 75, 2707–2720.
Saha, S., Raghava, G.P.S., 2007. Prediction of neurotoxins based on their function and source. In Silico Biol. 7, 0025.
Sanz, L., Gibbs, H.L., Mackessy, S.P., Calvete, J.J., 2006. Venom proteomes of closely related *Sistrurus* rattlesnakes with divergent diets. J. Proteome Res. 5, 2098–2112.
Schnoes, A.M., Brown, S.D., Dodevski, I., Babbitt, P.C., 2009. Annotation error in public databases, misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. 5 (12), e1000605. http://dx.doi.org/10.1371/journal.pcbi.1000605.
Sells, P.G., Ioannou, P., Theakston, R.D.G., 1998. A humane alternative to the measurement of the lethal effects, LD50, of non-neurotoxic venoms using hen's eggs. Toxicon 36, 985–991.
Siew, J.P., Khan, A.M., Tan, P.T., Koh, J.L., Seah, S.H., Koo, C.Y., Chai, S.C., Armugam, A., Brusic, V., Jeyaseelan, K., 2004. Systematic analysis of snake neurotoxins' functional classification using a data warehousing approach. Bioinformatics 20, 3466–3480.
Soogarun, S., Sangvanich, P., Chowbumroongkait, M., Jiemsup, S., Wiwanikit, V., Pradniwat, P., Palasuwan, A., Pawinwongchai, J., Chanprasert, S., Moungkote, T., 2008. Analysis of green pit viper, *Trimeresurus alborabris*. venom protein by LC/MS-MS. J. Biochem. Mol. Toxicol. 22, 225–229.
Sun, G.Y., Xu, J., Jensen, M.D., Simonyi, A., 2004. Phospholipase $A_2$ in the central nervous system, implications for neurodegenerative diseases. J. Lipid Res. 45, 205–213.
Tan, N.H., Tan, C.S., 1989. Fractionation of Sumatran pit viper, *Trimeresurus sumatranus sumatranus*, venom by DEAE-Sephacel ion exchange chromatography and some biological properties of the fractions. Toxicon 27, 697–702.
Tan, N.H., Tan, C.S., Khor, H.T., 1989. Isolation and characterization of the major phospholipase $A_2$ from the venom of *Trimeresurus purpureomaculatus*, shore pit viper. Int. J. Biochem. 21, 1421–1426.
Tan, P.T., Khan, A.M., Brusic, V., 2003. Bioinformatics for venom and toxin sciences. Brief. Bioinforma. 4, 53–62.
Tsai, I.H., Chen, Y.H., Wang, Y.M., Tu, A.T., 2003. Geographic variations, cloning and functional analyses of the venom acidic phospholipases $A_2$ of *Crotalus viridis viridis*. Arch. Biochem. Biophys. 411, 289–296.
Tsai, I.H., Wang, Y.M., Au, L.C., Ko, T.P., Chen, Y.H., Chu, Y.F., 2000. Phospholipases $A_2$ from *Calloselasma rhodostoma* venom gland. Cloning and sequencing of 10 of the cDNAs, three-dimensional modelling and chemical modification of the major isozyme. Eur. J. Biochem. 267, 6684–6691.
Wagstaff, S., Harrison, R., 2006. Venom gland EST analysis of the saw-scaled viper, *Echis ocellatus*, reveals novel alpha(9)beta(1) integrin-binding motifs in venom metalloproteinases and a new group of putative toxins, renin-like aspartic proteases. Gene 377, 21–32.
Wang, Y.M., Peng, H.F., Tsai, I.H., 2005. Unusual venom phospholipases $A_2$ of two primitive tree vipers *Trimeresurus puniceus* and *Trimeresurus borneensis*. FEBS J. 272, 3015–3025. http://dx.doi.org/10.1111/j.1742-4658.2005.04715.x.
Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview version 2, a multiple sequence alignment and analysis Workbench. Bioinformatics 25, 1189–1191. http://dx.doi.org/10.1093/bioinformatics/btp033.
Wei, J.F., Wei, X.L., Chen, Q.Y., Huang, T., Qiao, L.Y., Wang, W.Y., Xiong, Y.L., He, S.H., 2006. N49 phospholipase $A_2$, a unique subgroup of snake venom group II phospholipase $A_2$. Biochim. Biophys. Acta 1760, 462–471.
Wolf, M., Ruderisch, B., Dandekar, T., Müller, T., 2008. ProfdistS, (profile-) distance based phylogeny on sequence-structure alignments. Bioinformatics 24, 2401–2402.
Zhang, B., Liu, Q., Yin, W., Zhang, X., Huang, Y., Luo, Y., Qiu, P., Su, X., Yu, J., Hu, S., Yan, G., 2006. Transcriptome analysis of *Deinagkistrodon acutus* venomous gland focusing on cellular structure and functional aspects using expressed sequence tags. BMC Genomics 7, 152.
Zhou, X., Tan, T.C., Valiyaveettil, S., Go, M.L.R., Kini, R.M., Velazquez-Campoy, A., Sivaraman, J., 2008. Structural characterization of myotoxic ecarpholin S from *Echis carinatus* venom. Biophys. J. 95, 3366–3380.